KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 2 урок > Онлайн/офлайн метрики и backtest

Содержание урока

Онлайн метрики: что такое хорошая онлайн метрика, иерархия онлайн метрик, как перейти от онлайн к офлайн метрикам

Онлайн метрики и свойства хороших метрик

Онлайн метрики отражают взаимодействие пользователей с продуктом в реальном времени и помогают понять, как изменения в продукте влияют на поведение и восприятие пользователей. Онлайн метрики считают на результатах BackTest'a и

Ключевые характеристики хорошей онлайн метрики включают:

- **Чувствительность к малым изменениям**

  Чувствительная метрика способна обнаруживать даже незначительные изменения в поведении пользователей или производительности системы. Это означает, что даже малые улучшения или ухудшения могут быть зафиксированы и проанализированы. Например, если интернет-магазин внедряет новую систему рекомендаций, чувствительная метрика позволит оценить, привело ли это к увеличению среднего чека, даже если изменение составило всего 1-2%.

- **Надежность**

  Надежная метрика обеспечивает точность и повторяемость результатов. Она должна быть устойчива к случайным колебаниям и внешним помехам, таким как сезонность или временные тренды. Это гарантирует, что изменения в метрике действительно отражают результаты действий или изменений в продукте, а не случайные вариации.

- **Эффективность вычислений**

  Метрика считается эффективной, если ее можно рассчитать и проанализировать быстро, без значительных затрат ресурсов или времени. Эффективность метрики критически важна в динамичной бизнес-среде, где решения нужно принимать оперативно.

- **Способствует анализу и дебагу алгоритмов**

  Хорошая онлайн метрика не только показывает эффективность изменений, но и помогает идентифицировать проблемы в алгоритмах или процессах. Например, если после внедрения новой системы динамического ценообразования упала маржа, метрика должна помочь локализовать проблему — в выборе цен, в модели спроса и так далее.

- **Интерпретируемость**

  Метрика должна быть понятна для всех заинтересованных сторон, включая менеджмент, разработчиков, маркетологов и аналитиков. Каждый участник процесса понимает, почему выбрана эта метрика, как считается, в каких единицах измеряется, как это отражается на бизнес процессах. Это обеспечивает общее понимание целей и результатов, а также способствует более эффективному принятию решений на основе данных.

- **Инклюзивность**

  Метрика должна учитывать данные по всему спектру пользователей или операций, не исключая никакие группы. Это гарантирует, что улучшения для одной группы пользователей не ведут к ухудшению показателей для другой, обеспечивая справедливую и всестороннюю оценку изменений.

Примеры онлайн метрик включают абсолютные значения, как выручка, **маржа** (прибыль) и **заказы**, а также относительные показатели, как **коэффициент конверсии**, **относительная маржа** (когда хотим контролировать долю маржи в выручке).

Иерархия онлайн метрик и переход к офлайн метрикам

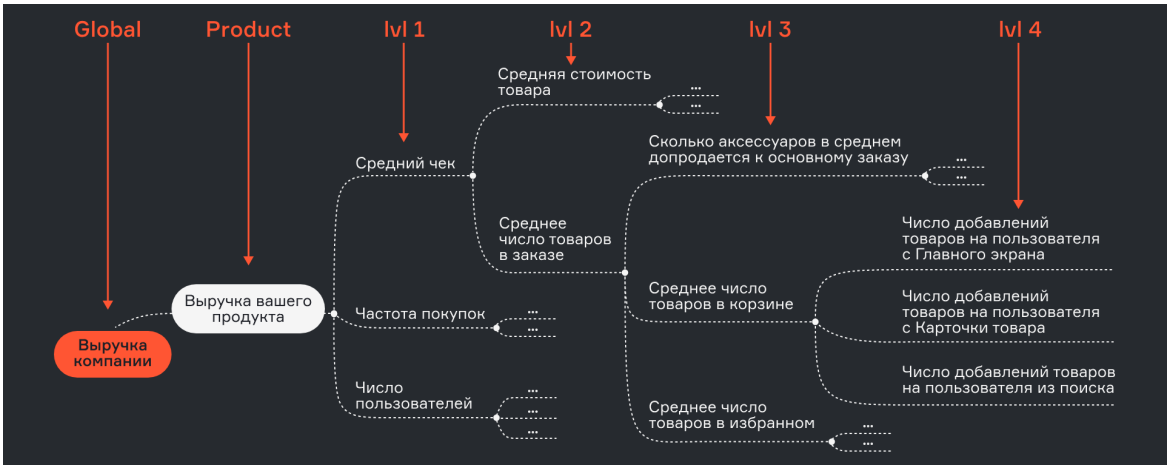
Структурирование метрик по иерархии позволяет анализировать и интерпретировать данные для принятия решений. Иерархия метрик обеспечивает логическую связь между различными уровнями метрик и их вкладом в общие бизнес-цели.

1. **Логическая структура**

   Центральное место в иерархии метрик занимает ключевая бизнес-цель, например, увеличение прибыли или рост числа активных пользователей. От этой цели отталкиваются все остальные метрики. Дочерние метрики должны быть напрямую связаны с родительскими, обеспечивая понимание того, как каждая метрика влияет на общую картину. Структурно метрики организованы как дерево, от одной родительской метрики (корня) до самых маленьких метрик (листочков). В этой аналогии также можно подметить, что дочерние метрики являются разложением родительской (корень+ствол → ветки → листья). Например, метрика "число новых пользователей" напрямую влияет на "общее число активных пользователей", которое, в свою очередь, влияет на "общую прибыль".

2. **Стабильность метрик**

   Важно, чтобы иерархия метрик была стабильной и не подвергалась частым изменениям. Это обеспечивает надежность данных и позволяет проводить сравнение показателей во времени. Метрики должны меняться только при значительных изменениях в бизнес-процессах или целях компании, что позволяет сохранять консистентность анализа и облегчает отслеживание долгосрочных тенденций.



Пример иерархии онлайн метрик

Преобразование онлайн метрик в офлайн метрики

Процесс превращения бизнес-метрик в метрики для анализа в офлайн режиме позволяет глубже понять взаимосвязи между действиями пользователей и их влиянием на бизнес. Например, бизнес-метрика "выручка", которая является онлайн метрикой, может быть преобразована в прокси-онлайн-метрику "средний чек", которая, в свою очередь, анализируется через прокси-оффлайн-метрику, такую как "отклонение от средней цены в категории", которая уже может быть преобразована в DS-метрики, например, MSE или RMSE. Это позволяет определить, как изменение цен влияет на покупательское поведение и общую выручку.

Офлайн метрики: что такое хорошая офлайн метрика, типы и примеры метрик, офлайн метрики в ML пайплайне

Офлайн метрики играют центральную роль в оценке и оптимизации моделей машинного обучения до их реализации в продуктовой среде. Эти метрики

позволяют провести глубокий анализ эффективности алгоритмов, основываясь на исторических данных, и определить их потенциальное влияние на ключевые бизнес-показатели. Рассмотрим более подробно основные типы офлайн метрик, некоторые примеры и их применение.

Абсолютные метрики

Абсолютные метрики предоставляют непосредственные числовые значения, которые могут быть использованы для оценки абсолютного эффекта от внедрения модели или изменений в бизнес-процессах.

- **WMAE (Weighted Mean Absolute Error, Средневзвешенная Абсолютная Ошибка):** Эта метрика измеряет среднее абсолютное отклонение прогнозируемых значений от фактических, с учетом веса каждого наблюдения. WMAE часто используется в задачах регрессии для оценки точности прогнозов модели, особенно когда некоторые ошибки имеют большее значение, чем другие.

$$WMAE = \frac{\sum_i^N (w_i |actual_i - predicted_i|)}{\sum_i^N w_i}$$

Относительные метрики

Относительные метрики сравнивают ошибку или эффективность модели относительно некоторой базовой линии или другого масштаба, что позволяет более наглядно оценить улучшение или ухудшение производительности.

- **WMAPE (Weighted Mean Absolute Percentage Error, Средневзвешенная Абсолютная Процентная Ошибка):** Эта метрика выражает ошибку как процент от фактических значений, что облегчает интерпретацию масштаба ошибок. WMAPE особенно полезна для сравнения эффективности моделей в различных масштабах данных.

$$WMAPE = \frac{\sum_i^N (w_i |actual_i - predicted_i|)}{\sum_i^N (w_i |actual_i|}$$

- **MASE (Mean Absolute Scaled Error, Средняя Абсолютная Масштабированная Ошибка):** Метрика сравнивает ошибку прогноза модели с ошибкой наивного прогноза, например, предсказания всегда среднего значения. MASE больше 1 указывает на то, что модель хуже наивного прогноза, а меньше 1 — лучше. Это позволяет оценить, насколько значимы улучшения, предоставляемые моделью.

$$MASE = \frac{MAE_{yourmodel}}{MAE_{naivemodel}}$$

Свойства хороших офлайн метрик

- **Скоррелирована с онлайн метрикой**

  Офлайн метрика должна связываться с влиянием на бизнес процесс, которое отражается в онлайн метриках. Например, если онлайн метрика отслеживает общий объем продаж или конверсию, то офлайн метрика должна предсказывать эти показатели с достаточной точностью. Такая корреляция позволяет оценить потенциальное влияние изменений цен до их реализации в реальных условиях, например, в backtest’е.

- **Интерпретируема и действенна**

  Метрика должна быть понятна всем участникам процесса, от разработчиков до бизнес-аналитиков и продуктовых менеджеров.

- **Не линейна**

  Метрика должна учитывать величину ошибки в прогнозах, поощряя более точные предсказания и штрафуя за значительные отклонения. Например, использование квадратичной функции потерь может помочь в достижении этого критерия, так как она накладывает более высокий штраф на большие ошибки.

- **Не чувствительна к выбросам**

  В условиях, когда могут происходить аномальные скачки спроса (например, из-за действий перекупщиков), метрика должна сохранять устойчивость и не допускать, чтобы эти выбросы существенно искажали общую картину.

- **Не смещена**

  Метрика должна одинаково оценивать степень и направление ошибки прогноза, будь то недопрогноз или перепрогноз.

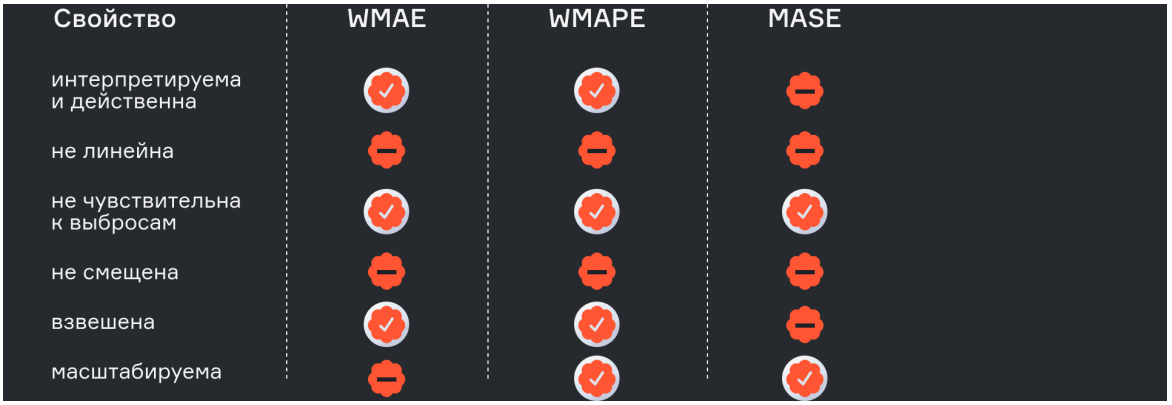
- **Взвешена**

  Важность различных категорий товаров в ассортименте может сильно различаться. Метрика должна учитывать эту важность, например, присваивая больший вес товарам высокого спроса или высокой маржинальности.

- **Масштабируема**

  Метрика должна адаптироваться к разным масштабам спроса и объемов продаж, обеспечивая адекватную оценку эффективности ценообразования как для мелких, так и для крупных категорий товаров.

Можно было бы предположить, что раз мы определили все необходимые свойства метрики - то наверняка придумали уже такую метрику, которая удовлетворяет всем этим критериям. Но, в реальности все несколько иначе, не существует такой метрики и всегда будет принимать решение о некотором трейдоффе между критериями.



Валидация пайплайна и офлайн метрики пайплайна

- Валидация модели спроса:

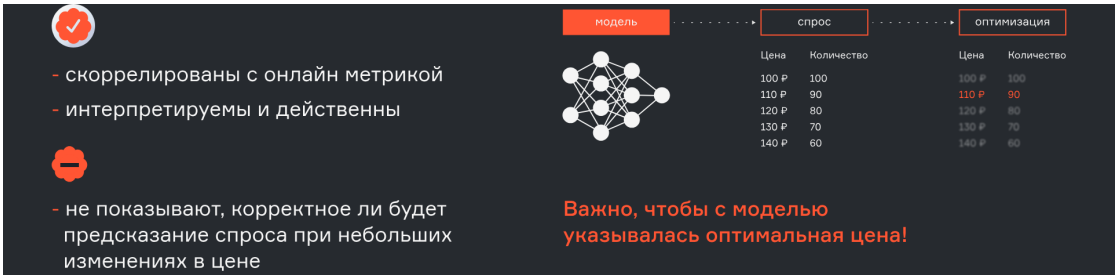
  Модель спроса должна отражать реальное поведение спроса. То есть, взяв произвольный товар, вы должны в подавляющем большинстве случаев наблюдать каноническую отрицательную кривую спроса, которая говорит о том, что при увеличении цены на товар спрос будет понижаться. При этом, вы можете дополнительно это провалидировать с помощью бизнеса, существуют категории товаров, которые заранее известны как “супер-эластичные” категории. Если при построении кривых спроса для этих категорий вы увидите что-то кроме эластичных кривых - что-то в модели спроса пошло не так. Сами модели вы валидируете с помощью офлайн метрик. Они показывают вам точность ваших прогнозов и позволяют прикинуть верно ли работает модель еще до построения кривых спроса.

- Валидация оптимизации:

  При решении задачи оптимизации с помощью метода Лагранжа (либо какого-то другого алгоритма оптимизации) вы получаете прогнозируемый аплифт целевой оптимизируемой метрики. Этот прогнозируемый аплифт поможет вам понять, адекватно ли решается задача оптимизации. Если вы на бектесте увидите аплифт 1%, в то время как ваш “оптимизатор” показал аплифт 10% - где-то есть баг и его нужно искать.

- Офлайн метрики пайплайна:

Офлайн метрики особенно важны при валидации моделей спроса, где необходимо точно оценить способность модели предсказывать будущий спрос на продукцию или услуги. Применение WMAE, WMAPE, MASE и других офлайн метрик позволяет не только оценить абсолютную и относительную ошибку прогнозов, но и сравнить эффективность модели с базовыми или наивными подходами, что критически важно для принятия решений о внедрении модели в производственную среду.



Важно понимать, что применение офлайн метрик в валидации модели спроса не означает, что вы будете точно знать о том, насколько сильно поднимется или опустится спрос на конкретный товар. Но вы будете знать, что он “поднимется” или “снизится” относительного некоторого начального состояния. И вот этот тезис верен для всех товаров, которые вы будете ценообразовывать. То есть, вы будете понимать, что условный спрос на телефоны сильнее реагирует на цену, чем спрос на мебельную фурнитуру.

Backtest

Офлайн метрики и метод backtest’a являются важными компонентами в процессе разработки и валидации ML алгоритмов, особенно в задачах, связанных с прогнозированием и оптимизацией. Эти инструменты позволяют оценить потенциальную эффективность модели в условиях, максимально приближенных к реальности, не подвергая риску текущие бизнес-процессы.

Применение офлайн метрик

Офлайн метрики применяются для оценки точности, эффективности и надежности моделей на исторических данных, то есть с помощью этих метрик вы сможете в backtest’е провалидировать ваши результаты.

Эти метрики позволяют:

1. **Оценить точность прогнозов:** Использование таких метрик, как WMAE и MASE, помогает оценить, насколько точно модель способна предсказывать реальные значения. Понять, насколько реальные у вас результаты.
2. **Сравнить модели:** Офлайн метрики позволяют провести сравнение между различными моделями или версиями одной модели, чтобы выбрать наиболее эффективную для достижения поставленной бизнес цели.

Метод Backtest

Backtest — это процесс тестирования модели на исторических данных для оценки, как бы модель вела себя в прошлом. Это ключевой шаг в оценке эффективности моделей, особенно в финансовой сфере, маркетинге и управлении запасами.

С помощью backtest'а можно:

1. **Использовать исторические данные:** Backtest позволяет "прогнать" модель на больших объемах исторических данных, что обеспечит понимание поведения модели в различных условиях.
2. **Оценить реальную эффективность:** Путем моделирования поведения модели в прошлом можно оценить ее эффективность и устойчивость к изменениям рыночных условий.
3. **Минимизировать потери компании:** С помощью backtest'а вы избежите запуска большинства провальных тестов → не потеряете время и деньги

Рекомендации по проведению Backtest

При проведении backtest важно учитывать следующие аспекты:

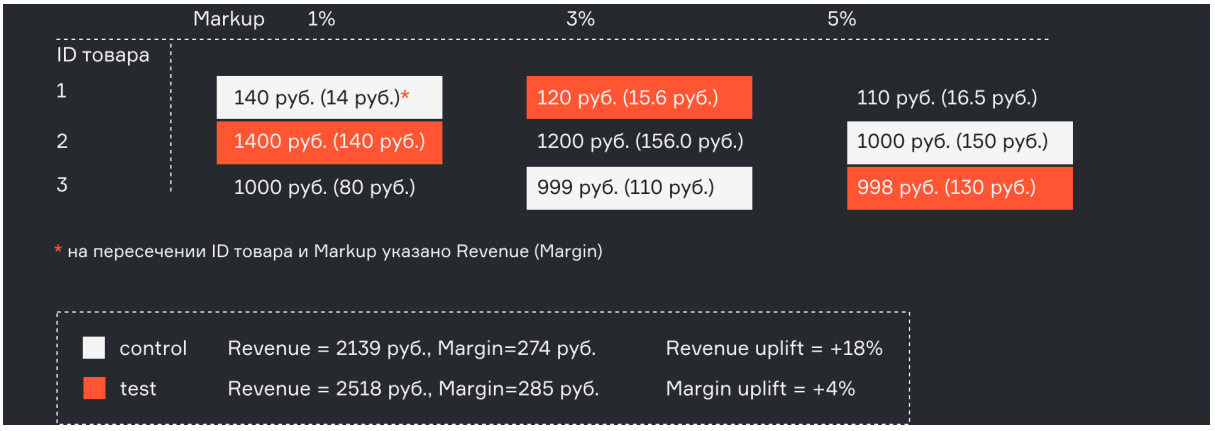
- **Отложенная выборка:** Используйте данные, которые не участвовали в обучении модели, для проведения теста. Это обеспечивает объективную оценку ее результатов. Размер отложенной выборки берите больше или равный длине планируемого теста.
- **Реалистичные условия:** Старайтесь максимально точно воспроизвести условия, в которых будет работать модель.
- **Сравнение с бейзлайном:** Всегда сравнивайте результаты модели с каким-либо бейзлайном или наивным прогнозом для ее оценки. Бейзлайном модели спроса может - плывущее среднее, бейзлайном модели динамического ценообразования - плоский маркап. (Плоский означает одинаковый на всех товарах)
- **Комплексный анализ метрик:** Используйте несколько офлайн метрик для оценки работы модели. Измеряйте метрики на крупных группах товаров.

Хорошо для этого подойдут категории товаров.

Правильно проведенный backtest в сочетании с анализом офлайн метрик позволяет повысить уверенность в том, что модель будет работать в реальных условиях, минимизируя потенциальные риски и повышая вероятность достижения ожидаемых бизнес-результатов.

В итоге, после проведения BackTest'а на руках у вас будет:

- Аплифт целевой метрики и вспомогательных метрик относительно контроля
- Назначенные маркапы на каждый товар
- Понимание, были ли такие маркапы в истории
- Ожидаемые целевые показатели



Результаты работы BackTest'а

Backtest. Практика

Сама имплементация алгоритма разобрана в практической части урока, ниже будут перечислены основные вещи, из которых состоит алгоритм, и о чем надо задумываться, когда вы будете алгоритм запускать и тестировать.

Как выглядит алгоритм?

1. Соединение исторических данных и предсказаний

2. Преобразование данных
    - Нормирование значений метрик на траффик (вы должны учить ваши модели на одинаковом траффике, если вы будете считать тоталы, без учета нормировки - вы будете наблюдать, что одна группа перформит кратно лучше другой)
    - Группировка по уровню, на котором хотим считать BackTest (например, по категории товара, комплектации, это нужно для того, чтобы увеличить покрытие по маркапам)
3. Поиск "ближайшего" маркапа для тестового и контрольного алгоритма (у вас не всегда будут все возможные маркапы для алгоритмов, поэтому мы добавляем некий tolerance к тому, что считаем “ближайшим” маркапом. Если eps=0.1, то 0.3 и 0.4, и 0.2 и 0.3 маркапы считаются ближайшими, но не 0.2 и 0.4)
4. Получение таблицы со статистикой

Советы из практики

- Смотрите пересечение маркапов на тесте и ваших алгоритмов
    - Если пересечение маленькое и аплифт низкий - это не означает, что ваш алгоритм плохой, это значит, что у вас недостаточно разнообразные данные для проведения Backtest'а ИЛИ можно перейти на уровень выше
- Не забывайте про честную оценку
    - Для Backtest'а используйте отложенную выборку, то есть ту, которая не участвовала в обучении алгоритма
    - Спрос подвержен сезонности, поэтому лучше проводить Backtest на разных периодах и брать период не меньше периода АБ-теста

Резюме урока

- **Онлайн метрики:**
    - Играют ключевую роль в анализе пользовательского поведения и эффективности изменений динамического ценообразования в реальном времени
    - Характеризуются чувствительностью к изменениям, надежностью, эффективностью вычислений, способностью к дебагу, интерпретируемостью и инклюзивностью
    - Включают абсолютные и относительные значения, такие как доход, маржа и коэффициент конверсии

    - Проверяйте качество онлайн метрик, насколько они адекватны для оценки вашего бизнес процесса
- **Иерархия онлайн метрик:**
    - Обеспечивает структурированный подход к анализу метрик, выявляя взаимосвязь между различными уровнями и их вкладом в общую бизнес-цель
    - Включает логическую структуру, стабильность метрик и помогает преобразовать онлайн метрики в офлайн DS метрики
    - Позволяет сформировать подход к оценке проекта динамического ценообразования
- **Офлайн метрики:**
    - Важны для оценки и оптимизации моделей машинного обучения перед их выпуском в реальный мир
    - Примеры: WMAE, WMAPE и MASE, используются для оценки точности, сравнения моделей
    - Нет идеальных метрик для оценки алгоритма. Всегда будет какой-то trade-off
    - Проверяйте качество ваших офлайн метрик. Если метрики будут плохими → плохая работа backtest’a → частый выпуск плохих моделей
- **Применение офлайн метрик и Backtest:**
    - Позволяет оценить потенциальную эффективность моделей на исторических данных
    - Необходимо использование отложенной выборки, реалистичных условий и сравнение с бейзлайном
- **Валидация пайплайна:**
    - Валидация модели спроса: Модель спроса должна отражать реальное поведение спроса. При разных ценах мы должны наблюдать разный спрос и величины спроса у разных товаров должны быть разными
    - Валидация оптимизации: При решении задачи оптимизации вы получаете какой-то аплифт оптимизируемой метрики. Этот аплифт должен соотноситься с тем, какой аплифт вы будете получать при проведении backtest’a
    - Офлайн метрики пайплайна: Офлайн метрики и backtest играют ключевую роль в валидации моделей спроса и модели оптимизации, позволяя оценить точность прогнозов и потенциальную эффективность модели

- **Рекомендации по разработке:**
    - Важно обеспечить понимание и согласованность метрик между командами бизнеса и ML
    - Необходимо договориться о ключевых метриках и убедиться в их взаимном понимании
    - Разработка архитектуры должна поддерживать удобство разработки и тестирования, включая декомпозицию на независимые части: подготовка данных, модель спроса, модель оптимизации и модуль офлайн тестирования